Advances in Data Analysis and Classification (2019) 13:773–795
https://doi.org/10.1007/s11634-018-0343-0

REGULAR ARTICLE

Variable selection in discriminant analysis for mixed continuous-binary variables and several groups

Alban Mbina Mbina[1] · **Guy Martial Nkiet**[1] · **Fulgence Eyi Obiang**[1]

Received: 12 March 2017 / Revised: 28 August 2018 / Accepted: 11 September 2018 /
Published online: 21 September 2018

Abstract
We propose a method for variable selection in discriminant analysis with mixed continuous and binary variables. This method is based on a criterion that permits to reduce the variable selection problem to a problem of estimating suitable permutation and dimensionality. Then, estimators for these parameters are proposed and the resulting method for selecting variables is shown to be consistent. A simulation study that permits to study several properties of the proposed approach and to compare it with an existing method is given, and an example on a real data set is provided.

Keywords Variable selection · Discriminant analysis · Classification · Mixed variables

Mathematics Subject Classification 62H30 · 62H12

1 Introduction

The problem of classifying an observation into one of several classes on the basis of data consisting of both continuous and categorical variables is an old problem that has been tackled under different forms in the literature. The earliest works in this field go back to Chang and Afifi (1979) and Krzanowski (1975) who used the location model introduced by Olkin and Tate (1961) to form a classification rule in the context of dis-

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11634-018-0343-0) contains supplementary material, which is available to authorized users.

✉ Guy Martial Nkiet
gnkiet@hotmail.com

Alban Mbina Mbina
albanmbinambina@yahoo.fr

Fulgence Eyi Obiang
feyiobiang@yahoo.fr

[1] Laboratoire URMI, Université des Sciences et Techniques de Masuku, BP 943, Franceville, Gabon

criminant analysis involving two groups. More recent work has focused on defining distance measures between populations or making inference on them (e.g., Krzanowski 1983, 1984; Bar-Hen and Daudin 1995; Bedrick et al. 2000; de Leon and Carriere 2005). One of the most important problem in the context described above is the problem of selecting the appropriate categorical and/or continuous variables to use for discrimination. Indeed, it is well recognized that using fewer variables improve classification performance and permits to avoid estimation problems (e.g. McLachlan 1992; Mahat et al. 2007). There are several works dealing with this problem, mainly in the context of a location model. Some of these works are based on the use of distances between populations for determining the most predictive variables (Krzanowski 1983; Daudin 1986; Bar-Hen and Daudin 1995; Daudin and Bar-Hen 1999). Krusinska (1989a, b, 1990) used methods based on the percentage of misclassification, Hotelling's T^2 and graphical models. More recently, Mahat et al. (2007) proposed a method based on distance between groups as measured by smoothed Kullback–Leiler divergence. All these works consider the case of two groups and, to the best of our knowledge, the case of more than two groups have not yet been considered for variable selection purpose. So, it is of great interest to introduce a method that can be used when the number of groups is greater than two. Such an approach have been proposed recently in Nkiet (2012) for the case of continuous variables only. It is based on a criterion that permits to characterize the set of variables that are appropriate for discrimination by means of two parameters, so that the variable selection problem reduces to that of estimating these parameters.

In this paper, we extend the approach of Nkiet (2012) to the case of mixed variables. The resulting method has two advantages; first, it can be used when the number of groups is greater than two, and secondly it just requires that the random vector consisting of the continuous variables has finite fourth order moment. No assumption on the distribution of this random vector is needed and, therefore, we do not suppose that the location model holds. In Sect. 2, we introduce a criterion by means of which the set of variables to be estimated is characterized by means of suitable permutation and dimensionality. Then, estimating this criterion is tackled in Sect. 3. More precisely, empirical estimators as well as non-parametric smoothing procedure are used for defining an estimator of the criterion. In the first case, we obtain properties of the resulting estimator that permits to obtain its asymptotic distribution. Section 4 is devoted to the definition of our proposal for variable selection. Consistency of the method, when empirical estimators are used, is then proved. Section 5 is devoted to the presentation of numerical experiments made in order to study several properties of the proposal and to compare it with an existing method. The first issue that is adressed concerns the impact of choosing penalty functions that are involved in our procedure, and that of the type of estimators that is used. The results reveal low impact on the performance of the proposed method, and that in case of very small cell incidences it is preferable to use smoothed estimators. Since this method depends on two real parameters, it is of interest to study their influence on its performance and, consequently, to define a strategy that permits to choose optimal values for them; we propose a method based on leave-one-out cross validation for obtaining these optimal values. When using this approach, the obtained results show that the proposal is competitive with that of Mahat et al. (2007). A real data example is given in Sect. 6.

2 Statement of the problem

Letting $(\Omega, \mathcal{A}, P)$ be a probability space, we consider random vectors $X = \left(X^{(1)}, \ldots, X^{(p)}\right)^T$ and $Y = \left(Y^{(1)}, \ldots, Y^{(d)}\right)^T$ defined on this probability space and with values in $\mathbb{R}^p$ and $\{0, 1\}^d$ respectively. The r.v. X consists of continuous random variables whereas Y consists of binary random variables. As usual, Y may be associated to a multinomial random variable by considering $U = 1 + \sum_{j=1}^{d} Y^{(j)} 2^{j-1}$ which has values in $\{1, \ldots, M\}$, where $M = 2^d$. In fact, Y is essentially the binary representation of U, and U provides a labeling of the underlying d-dimensional contingency table. Suppose that the observations of (X, Y) come from q groups $\pi_1, \ldots \pi_q$ (with $q \geq 2$) characterized by a random variable Z with values in $\{1, \ldots, q\}$; this means that (X, Y) belongs to π_ℓ if, and only if, one has $Z = \ell$. Such framework has been considered in the literature for classification purposes. Indeed, for the case of two groups, that is when $q = 2$, Krzanowski (1975) proposed a classification rule based on a normal distribution location model where, conditionally on $U = m$ and $Z = \ell$, the vector X has the p-variate distribution $N(\mu_{m\ell}, \Sigma)$. This rule allocates a future observation (x, y) of (X, Y) to π_1 if

$$(\mu_{m1} - \mu_{m2})^T \Sigma^{-1} \left(x - \frac{1}{2}(\mu_{m1} + \mu_{m2}) \right) \geq \log \left(\frac{p_{m2}}{p_{m1}} \right) + \log(\gamma), \qquad (1)$$

where $m = 1 + \sum_{j=1}^{d} y^{(j)} 2^{j-1}$, $p_{m\ell} = P(U = m | Z = \ell)$ and γ is a constant that depends on costs due to missclassification and prior probabilities for the two groups. The case where $q > 2$ was considered by de Leon et al. (2011) in the context of general mixed-data models. In this case, the optimum rule classifies an observation (x, y) into the class π_{ℓ^*} if

$$\delta_m^{(\ell^*)}(x, y) = \max_{\ell=1,\ldots,q} \delta_m^{(\ell)}(x, y) \qquad (2)$$

where

$$\delta_m^{(\ell)}(x, y) = (\mu_{m\ell})^T \Sigma^{-1} x - \frac{1}{2}(\mu_{m\ell})^T \Sigma^{-1} \mu_{m\ell} + \log(p_{m\ell}) + \log(\tau_\ell), \qquad (3)$$

with $\tau_\ell = P(Z = \ell)$. As it can be seen, these rules involve observations of all the variables $X^{(j)}$ in X. Nevertheless, as it is well recognized (see, e.g., McLachlan 1992; Mahat et al. 2007), using fewer variables may improve classification performance. So, it is of real interest to perform selection of the $X^{(j)}$'s from a sample of (X, Y, Z). For doing that, we extend an approach proposed by Nkiet (2012) for the case of continuous variables to the case of mixed continuous-binary variables where selection is from the continuous variables only. This approach first consists in introducing a criterion by means of which the set of variables that are adequate for discrimination is characterized. For any $m \in \{1, \ldots, M\}$ we put

$$p_m = P(U = m), \;\; \mu_m = \mathbb{E}(X | U = m), \;\; \mu_{\ell,m} = \mathbb{E}(X | Z = \ell, U = m)$$

and, assuming that $\mathbb{E}\left(\|X\|^2\right) < +\infty$ where $\|\cdot\|$ denotes the usual Euclidean norm of $\mathbb{R}^p$, we consider the covariance matrix of X conditionally on $U = m$ given by

$$V_m = \mathbb{E}\left((X - \mu_m)(X - \mu_m)^T | U = m\right),$$

wher a^T denotes the transpose of a. Througout this paper we assume that the matrix V_m is invertible. Let us represent the set of continuous variables by the set $I = \{1, \ldots, p\}$ and, for any subset $K := \{i_1, \ldots, i_k\}$ of I, consider the $k \times p$ matrix defined by:

$$A_K = \begin{pmatrix} a_{11}^{(K)} & a_{12}^{(K)} & \cdots & a_{1p}^{(K)} \\ a_{21}^{(K)} & a_{22}^{(K)} & \cdots & a_{2p}^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1}^{(K)} & a_{k2}^{(K)} & \cdots & a_{kp}^{(K)} \end{pmatrix}$$

where

$$a_{lj}^{(K)} = \begin{cases} 1 & \text{if } j = i_l \\ 0 & \text{if } j \neq i_l \end{cases}, \ 1 \leq l \leq k, \ 1 \leq j \leq p.$$

This matrix selects from x in $\mathbb{R}^p$ the components contained in the set K; more precisely, A_K transforms any vector $x = (x_1, \ldots, x_p)^T$ in $\mathbb{R}^p$ to the vector $A_K x$ in $\mathbb{R}^k$ whose components are the components x_i of x such that $i \in K$. Then putting

$$p_{\ell|m} = P(Z = \ell | U = m) \quad \text{and} \quad Q_{K|m} = A_K^T \left(A_K V_m A_K^T\right)^{-1} A_K \tag{4}$$

we introduce

$$\xi_{K|m} = \sum_{\ell=1}^{q} p_{\ell|m}^2 \| \left(\mathbb{I}_p - V_m Q_{K|m}\right)\left(\mu_{\ell,m} - \mu_m\right) \|^2, \tag{5}$$

where $\mathbb{I}_p$ is the $p \times p$ identity matrix. In fact, $\xi_{K|m}$ is the criterion introduced in Nkiet (2012), conditionally on $U = m$. It quantifies the loss of information resulting from selection of the variables of X belonging to K provided that the event $\{U = m\}$ is realized (see Theorem 2.1 in Nkiet 2012). From this, we look for a criterion that quantifies the loss of information resulting from the above variable selection but without taking into account the value that will be taken by U since this value is unknown a priori. This leads to consider the criterion

$$\xi_K = \sum_{m=1}^{M} p_m^2 \xi_{K|m} \tag{6}$$

by means of which we will characterize the subset I_0 of variables that do not make any contribution for the discrimination between the q groups, that is variables of which the related components of the vector $\Sigma^{-1}\mu_{m\ell}$ that arises in the discriminant function (3) are constant as ℓ varies in $\{1, \ldots, q\}$, for any $m \in \{1, \ldots, M\}$. Therefore,

$$I_0 = \bigcap_{m=1}^{M} I_{0,m}$$

where $I_{0,m}$ denotes the subset of continuous variables whose related components of $\Sigma^{-1}\mu_{m\ell}$ are constant for $\ell \in \{1, \ldots, q\}$. Then, our problem reduces to the problem of estimating the subset I_1 given by

$$I_1 = I - I_0 = \bigcup_{m=1}^{M} I_{1,m}$$

where $I_{1,m} = I - I_{0,m}$. An explicit expression of $I_{1,m}$ can be obtained by using results from McKay (1977) (see also Fujikoshi 1982, 1985). Indeed, let $\lambda_{1,m} \geq \lambda_{2,m} \geq \cdots \geq \lambda_{p,m}$ denote the eigenvalues of $T_m = V_m^{-1} B_m$ where B_m is the between groups covariance matrix conditionally on $U = m$ given by

$$B_m = \sum_{\ell=1}^{q} p_{\ell|m}(\mu_{\ell,m} - \mu_m)(\mu_{\ell,m} - \mu_m)^T,$$

and let $\upsilon_i^m = (\upsilon_{i1}^m, \ldots, \upsilon_{ip}^m)^T$ $(i = 1, \ldots, p)$ be an eigenvector of T_m associated with $\lambda_{i,m}$, then (see McKay 1977; Fujikoshi 1982),

$$I_{1,m} = \left\{k \in I | \exists i \in \{1, \ldots, r_m\}, \upsilon_{ik}^m \neq 0\right\},$$

where r_m denotes the rank of T_m.
Now, we give a characterization of I_1 by means of the criterion (6). For doing that, we consider the following assumption:
$(\mathcal{A}_1)$: For all $m \in \{1, \ldots, M\}$, $p_m > 0$.
Then, we have:

Proposition 1 *We assume that* $(\mathcal{A}_1)$ *holds. Then, for* $K \subset I$ *we have* $\xi_K = 0$ *if and only if* $I_1 \subset K$.

Proof For any fixed $m \in \{1, \ldots, M\}$, we denote by $P^{\{U=m\}}$ the conditional probability to the event $\{U = m\}$. Then applying Theorem 2.1 of Nkiet (2012) with the probability space $\left(\Omega, \mathcal{A}, P^{\{U=m\}}\right)$, we obtain the equivalence: $\xi_{K|m} = 0 \Leftrightarrow I_{1,m} \subset K$. Thus:

$$\begin{aligned} \xi_K = 0 &\Leftrightarrow \sum_{m=1}^{M} p_m^2 \xi_{K|m} = 0 \\ &\Leftrightarrow \forall m \in \{1, \ldots, M\}, \xi_{K|m} = 0 \end{aligned}$$

$$\Leftrightarrow \forall m \in \{1, \ldots, M\}, I_{1,m} \subset K$$
$$\Leftrightarrow \bigcup_{m=1}^{M} I_{1,m} \subset K.$$

□

From this proposition, it is easily seen that, putting $K_i = I - \{i\}$, one has the equivalence: $\xi_{K_i} > 0 \Leftrightarrow i \in I_1$. Now, let $\sigma = (\sigma(1), \ldots, \sigma(p))$ be the permutation of I such that:

(a) $\xi_{K_{\sigma(1)}} \geq \xi_{K_{\sigma(2)}} \geq \cdots \geq \xi_{K_{\sigma(p)}}$;
(b) $\xi_{K_{\sigma(i)}} = \xi_{K_{\sigma(j)}}$ and $i < j$ imply $\sigma(i) < \sigma(j)$.

Remark 1 This just means that the ξ_{K_i}'s are ranked in nonincreasing order so that the ex aequos are ranked in increasing order of the corresponding indices. Then, $\sigma(i)$ is the index corresponding to the i-th largest element.

Since I_1 is a non-empty set, there exists an integer $s \in I$ which is equal to p when $I_1 = I$, and satisfying

$$\xi_{K_{\sigma(1)}} \geq \cdots \geq \xi_{K_{\sigma(s)}} > \xi_{K_{\sigma(s+1)}} = \cdots = \xi_{K_{\sigma(p)}} = 0$$

when $I_1 \neq I$. The integer s will be the number of selected variables, and:

$$I_1 = \{\sigma(i);\ 1 \leq i \leq s\}. \tag{7}$$

Therefore, estimating I_1 reduces to estimating the two parameters σ and s. For doing that, we first need to consider an estimator of the criterion given in (6).

3 Estimating the criterion

Let $\{(X_i, Y_i, Z_i)\}_{1 \leq i \leq n}$ be an i.i.d. sample of (X, Y, Z) with

$$X_i = \left(X_i^{(1)}, \ldots, X_i^{(p)}\right)^T \quad \text{and} \quad Y_i = \left(Y_i^{(1)}, \ldots, Y_i^{(d)}\right)^T;$$

we put $U_i = 1 + \sum_{j=1}^{d} Y_i^{(j)} 2^{j-1}$. In this section, we define estimators for the criterion given in (6) by estimating the parameters involved in its definition. First, empirical estimators are introduced and properties of the resulting estimator of the criterion are given, and secondly we consider estimators obtained by using non-parametric smoothing procedures as in Mahat et al. (2007).

3.1 Empirical estimators

Putting

$$N_m^{(n)} = \sum_{i=1}^{n} \mathbf{1}_{\{U_i=m\}} \text{ and } N_{\ell,m}^{(n)} = \sum_{i=1}^{n} \mathbf{1}_{\{Z_i=\ell, U_i=m\}},$$

we estimate p_m, $p_{\ell|m}$, μ_m, $\mu_{\ell,m}$ and V_m respectively by:

$$\widehat{p}_m^{(n)} = \frac{N_m^{(n)}}{n}, \quad \widehat{p}_{\ell|m}^{(n)} = \frac{N_{\ell,m}^{(n)}}{N_m^{(n)}}, \quad \widehat{\mu}_m^{(n)} = \frac{1}{N_m^{(n)}} \sum_{i=1}^{n} \mathbf{1}_{\{U_i=m\}} X_i,$$

$$\widehat{\mu}_{\ell,m}^{(n)} = \frac{1}{N_{\ell,m}^{(n)}} \sum_{i=1}^{n} \mathbf{1}_{\{Z_i=\ell, U_i=m\}} X_i$$

and

$$\widehat{V}_m^{(n)} = \frac{1}{N_m^{(n)}} \sum_{i=1}^{n} \mathbf{1}_{\{U_i=m\}} \left(X_i - \widehat{\mu}_m^{(n)}\right) \left(X_i - \widehat{\mu}_m^{(n)}\right)^T.$$

Then, considering

$$\widehat{Q}_{K|m}^{(n)} = A_K^T \left(A_K \widehat{V}_m^{(n)} A_K^T\right)^{-1} A_K$$

and

$$\widehat{\xi}_{K|m}^{(n)} = \sum_{\ell=1}^{q} \left(\widehat{p}_{\ell|m}^{(n)}\right)^2 \| \left(\mathbb{I}_p - \widehat{V}_m^{(n)} \widehat{Q}_{K|m}^{(n)}\right) \left(\widehat{\mu}_{\ell,m}^{(n)} - \widehat{\mu}_m^{(n)}\right) \|^2,$$

we take as estimator of ξ_K the random variable $\widehat{\xi}_K^{(n)}$ defined by:

$$\widehat{\xi}_K^{(n)} = \sum_{m=1}^{M} \left(\widehat{p}_m^{(n)}\right)^2 \widehat{\xi}_{K|m}^{(n)}. \tag{8}$$

Now, we will derive a result which establishes strong consistency for $\widehat{\xi}_K^{(n)}$ and will be useful for determining its asymptotic distribution and consistency of the proposed method for selecting variables for n approaching $+\infty$. Let us consider the random matrices

$$\mathcal{X} = \begin{pmatrix} \mathbf{1}_{\{Z=1,U=1\}} X & \dots & \mathbf{1}_{\{Z=1,U=M\}} X \\ \mathbf{1}_{\{Z=2,U=1\}} X & \dots & \mathbf{1}_{\{Z=2,U=M\}} X \\ \vdots & \ddots & \vdots \\ \mathbf{1}_{\{Z=q,U=1\}} X & \dots & \mathbf{1}_{\{Z=q,U=M\}} X \end{pmatrix},$$

$$\mathcal{X}_i = \begin{pmatrix} \mathbf{1}_{\{Z_i=1,U_i=1\}}X_i & \dots & \mathbf{1}_{\{Z_i=1,U_i=M\}}X_i \\ \mathbf{1}_{\{Z_i=2,U_i=1\}}X_i & \dots & \mathbf{1}_{\{Z_i=2,U_i=M\}}X_i \\ \vdots & \ddots & \vdots \\ \mathbf{1}_{\{Z_i=q,U_i=1\}}X_i & \dots & \mathbf{1}_{\{Z_i=q,U_i=M\}}X_i \end{pmatrix}$$

with values in the space $\mathcal{M}_{qp,M}(\mathbb{R})$ of $pq \times M$ matrices. We also introduce the random vectors

$$\mathcal{Y} = \begin{pmatrix} \mathbf{1}_{\{U=1\}}X \\ \mathbf{1}_{\{U=2\}}X \\ \vdots \\ \mathbf{1}_{\{U=M\}}X \end{pmatrix}, \ \mathcal{Y}_i = \begin{pmatrix} \mathbf{1}_{\{U_i=1\}}X_i \\ \mathbf{1}_{\{U_i=2\}}X_i \\ \vdots \\ \mathbf{1}_{\{U_i=M\}}X_i \end{pmatrix}, \ \mathcal{Z} = \begin{pmatrix} \mathbf{1}_{\{U=1\}} \\ \mathbf{1}_{\{U=2\}} \\ \vdots \\ \mathbf{1}_{\{U=M\}} \end{pmatrix}, \ \mathcal{Z}_i = \begin{pmatrix} \mathbf{1}_{\{U_i=1\}} \\ \mathbf{1}_{\{U_i=2\}} \\ \vdots \\ \mathbf{1}_{\{U_i=M\}} \end{pmatrix},$$

and the random matrices

$$\mathcal{U} = \begin{pmatrix} \mathbf{1}_{\{Z=1,U=1\}} & \dots & \mathbf{1}_{\{Z=1,U=M\}} \\ \mathbf{1}_{\{Z=2,U=1\}} & \dots & \mathbf{1}_{\{Z=2,U=M\}} \\ \vdots & \ddots & \vdots \\ \mathbf{1}_{\{Z=q,U=1\}} & \dots & \mathbf{1}_{\{Z=q,U=M\}} \end{pmatrix}, \ \mathcal{U}_i = \begin{pmatrix} \mathbf{1}_{\{Z_i=1,U_i=1\}} & \dots & \mathbf{1}_{\{Z_i=1,U_i=M\}} \\ \mathbf{1}_{\{Z_i=2,U_i=1\}} & \dots & \mathbf{1}_{\{Z_i=2,U_i=M\}} \\ \vdots & \ddots & \vdots \\ \mathbf{1}_{\{Z_i=q,U_i=1\}} & \dots & \mathbf{1}_{\{Z_i=q,U_i=M\}} \end{pmatrix},$$

and

$$\mathcal{V} = \begin{pmatrix} \mathbf{1}_{\{U=1\}}XX^T \\ \mathbf{1}_{\{U=2\}}XX^T \\ \vdots \\ \mathbf{1}_{\{U=M\}}XX^T \end{pmatrix}, \ \mathcal{V}_i = \begin{pmatrix} \mathbf{1}_{\{U_i=1\}}X_iX_i^T \\ \mathbf{1}_{\{U_i=2\}}X_iX_i^T \\ \vdots \\ \mathbf{1}_{\{U_i=M\}}X_iX_i^T \end{pmatrix}.$$

Further, we consider the random matrices

$$\mathcal{W} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{U}, \mathcal{V}), \ \mathcal{W}_i = (\mathcal{X}_i, \mathcal{Y}_i, \mathcal{Z}_i, \mathcal{U}_i, \mathcal{V}_i)$$

with values in

$$\mathcal{E} = \mathcal{M}_{qp,M}(\mathbb{R}) \times \mathbb{R}^{pM} \times \mathbb{R}^M \times \mathcal{M}_{q,M}(\mathbb{R}) \times \mathcal{M}_{pM,p}(\mathbb{R});$$

then, we put

$$\widehat{W}^{(n)} = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{W}_i - \mathbb{E}(\mathcal{W})\right). \tag{9}$$

Note that, for any $(a, b, c, d, e) \in \mathcal{E}$, we can write:

$$a = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,M} \\ a_{2,1} & a_{2,2} & \dots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q,1} & a_{q,2} & \dots & a_{q,M} \end{pmatrix}, \text{ where } a_{\ell,m} \in \mathbb{R}^p, \ 1 \le \ell \le q, \ 1 \le m \le M,$$

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{pmatrix}, \text{ where } b_m \in \mathbb{R}^p, 1 \le m \le M,$$

$$c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{pmatrix}, \text{ where } c_m \in \mathbb{R}, \ 1 \le m \le M,$$

$$d = \begin{pmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,M} \\ d_{2,1} & d_{2,2} & \dots & d_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{q,1} & d_{q,2} & \dots & d_{q,M} \end{pmatrix}, \text{ where } c_{\ell,m} \in \mathbb{R}, \ 1 \le \ell \le q, \ 1 \le m \le M,$$

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_M \end{pmatrix}, \text{ where } e_m \in \mathcal{M}_{p,p}(\mathbb{R}), 1 \le m \le M.$$

We introduce the projectors

$$\begin{aligned} \pi_1^{\ell m} &: (a,b,c,d,e) \in \mathcal{E} \mapsto a_{\ell,m} \in \mathbb{R}^p, \\ \pi_2^{m} &: (a,b,c,d,e) \in \mathcal{E} \mapsto b_{m} \in \mathbb{R}^p, \\ \pi_3^{m} &: (a,b,c,d,e) \in \mathcal{E} \mapsto c_{m} \in \mathbb{R}, \\ \pi_4^{\ell m} &: (a,b,c,d,e) \in \mathcal{E} \mapsto d_{\ell,m} \in \mathbb{R}, \\ \pi_5^{m} &: (a,b,c,d,e) \in \mathcal{E} \mapsto e_{m} \in \mathcal{M}_{p,p}(\mathbb{R}), \end{aligned}$$

the vector

$$\Delta_{\ell,K|m} = \left(\mathbb{I}_p - V_m Q_{K|m}\right)\left(\mu_{\ell,m} - \mu_m\right)$$

and the maps $\Lambda_{K|m}$, $\Phi_{\ell,K|m}$ and $\Psi_{\ell,K|m}$ defined on $\mathcal{E}$ by

$$\begin{aligned} \Lambda_{K|m}(S) &= 2 p_m \pi_3^m(S)\, \xi_{K|m}, \\ \Phi_{\ell,K|m}(S) &= p_m^{-1}\left(\pi_4^{\ell m}(S) - \pi_3^m(S)\, p_{\ell|m}\right) \|\Delta_{\ell,K|m}\|, \end{aligned}$$

and

$$\begin{aligned} \Psi_{\ell,K|m}(S) = p_m^{-1}\left(\mathbb{I}_p - V_m Q_{K,m}\right) \Big[& p_{\ell|m}^{-1}\left(\pi_1^{\ell m}(S) - \pi_4^{\ell m}(S)\mu_{\ell,m}\right) - \pi_2^m(S) + \pi_3^m(S)\mu_m \\ & - \left(\pi_5^m(S) - \pi_3^m(S)\left(V_m + \mu_m \mu_m^T\right)\right) Q_{K|m}\left(\mu_{\ell,m} - \mu_m\right) \end{aligned}$$

$$+\left(\mu_m\left(\pi_2^m(S)-\pi_3^m(S)\mu_m\right)^T\right)Q_{K|m}\left(\mu_{\ell,m}-\mu_m\right)$$
$$+\left(\left(\pi_2^m(S)-\pi_3^m(S)\mu_m\right)\mu_m^T\right)Q_{K|m}\left(\mu_{\ell,m}-\mu_m\right)\Big].$$

Then, we can formulate the following theorem that asserts the consistency of $\widehat{\xi}_K^{(n)}$ for n approaching $+\infty$. Moreover, we provide an asymptotic approximation for $\widehat{\xi}_K^{(n)}$ that will be useful for deriving its asymptotic distribution. Let us introduce the following assumptions:
$(\mathcal{A}_2)$: For all $(\ell, m) \in \{1, \ldots, q\} \times \{1, \ldots, M\}$, $p_{\ell|m} > 0$;
$(\mathcal{A}_3)$: $\mathbb{E}\left(\|X\|^2\right) < +\infty$.
Then, we have:

Theorem 1 *We assume that the asumptions* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_3)$ *hold. Then, for any subset* K *of* I *we have:*

(i) $\widehat{\xi}_K^{(n)}$ *converges almost surely to* ξ_K *as* $n \to +\infty$.
(ii)

$$n\widehat{\xi}_K^{(n)} = \sum_{m=1}^{M}\sqrt{n}\widehat{\Lambda}_{K|m}^{(n)}(\widehat{W}^{(n)}) + \sum_{m=1}^{M}\sum_{\ell=1}^{q}\Big(p_m\widehat{\Phi}_{\ell,K|m}^{(n)}\left(\widehat{W}^{(n)}\right) + p_m p_{\ell|m}\|\widehat{\Psi}_{\ell,K|m}^{(n)}\left(\widehat{W}^{(n)}\right) + \sqrt{n}\Delta_{\ell,K|m}\|\Big)^2$$

where $\left(\widehat{\Lambda}_{K|m}^{(n)}\right)_{n\in\mathbb{N}^*}$, $\left(\widehat{\Phi}_{\ell,K|m}^{(n)}\right)_{n\in\mathbb{N}^*}$ *and* $\left(\widehat{\Psi}_{\ell,K|m}^{(n)}\right)_{n\in\mathbb{N}^*}$ *are sequences of random operators that converge almost surely uniformly to* $\Lambda_{K|m}$, $\Phi_{\ell,K|m}$ *and* $\Psi_{\ell,K|m}$ *respectively.*

3.2 Non-parametric smoothing procedure

As it is well known, empirical estimators could be not suitable, because many cell entries of the d-variate contingency table corresponding to the d-categorical variables could be very small, so that corresponding estimates could be poor or non-existent (see Aspakourov and Krzanowski 2000). To overcome these problems, smoothed non-parametric estimators (of probabilities, means and covariance matrices) introduced in Aspakourov and Krzanowski (2000) and Mahat et al. (2007) can be used in order to estimate the criterion (6). Denote by D the dissimilarity defined on $\{1, \ldots, M\}^2$ by

$$D(m, k) = \|\mathbf{y}_m - \mathbf{y}_k\|^2,$$

where, for any $k \in \{1, \ldots, M\}$, $\mathbf{y}_k = \left(\mathbf{y}_k^{(1)}, \ldots, \mathbf{y}_k^{(d)}\right)^T \in \{0, 1\}^d$ is the vector of binary variables satisfying $1 + \sum_{j=1}^{d}\mathbf{y}_k^{(j)}2^{j-1} = k$. Then, given a smoothing parameter $\lambda \in]0, 1[$, we consider the weights $w(m, k) = \lambda^{D(m,k)}$ and estimate p_m, $p_{\ell|m}$, μ_m, $\mu_{\ell,m}$ and V_m respectively by:

$$\widetilde{p}_m^{(n)} = \frac{\sum_{j=1}^M w(m,j)N_j^{(n)}}{\sum_{k=1}^M \sum_{j=1}^M w(k,j)N_j^{(n)}}, \quad \widetilde{p}_{\ell|m}^{(n)} = \frac{\sum_{j=1}^M w(m,j)N_{\ell,j}^{(n)}}{\widetilde{p}_m^{(n)} \sum_{k=1}^M \sum_{j=1}^M w(k,j)N_{\ell,j}^{(n)}},$$

$$\widetilde{\mu}_m^{(n)} = \left\{ \sum_{j=1}^M w(m,j)N_j^{(n)} \right\}^{-1} \sum_{j=1}^M \left\{ w(m,j) \sum_{i=1}^n \mathbf{1}_{\{U_i=j\}} X_i \right\},$$

$$\widetilde{\mu}_{\ell,m}^{(n)} = \left\{ \sum_{j=1}^M w(m,j)N_{\ell,j}^{(n)} \right\}^{-1} \sum_{j=1}^M \left\{ w(m,j) \sum_{n=1}^n \mathbf{1}_{\{Z_i=\ell,U_i=j\}} X_i \right\}$$

and

$$\widetilde{V}_m^{(n)} = \left\{ \sum_{j=1}^M w(m,j)N_j^{(n)} \right\}^{-1} \sum_{j=1}^M \left\{ w(m,j) \sum_{i=1}^n \mathbf{1}_{\{U_i=j\}} (X_i - \widetilde{\mu}_m^{(n)})(X_i - \widetilde{\mu}_m^{(n)})^T \right\}.$$

Then, we obtain an estimator $\widetilde{\xi}_K^{(n)}$ of the criterion by replacing in (4), (5) and (6) the parameters p_m, $p_{\ell|m}$, μ_m, $\mu_{\ell,m}$ and V_m by their estimators given above. Note that the smoothing parameter λ must be obtained before parameters can be estimated. Mahat et al. (2007) suggested to choose a value of λ that gives good performance of a classification rule and proposed to obtain it by minimizing Brier score (e.g. Hand 1997, p. 101). No consistency result has been given for the aforementioned estimators.

4 Selection of variables

From Eq. (7) it is seen that estimation of I_1 reduces to the determination of the permutation $\sigma = (\sigma(1), \ldots, \sigma(p))$ and the number of selected variables s. In this section, estimators of σ and s are proposed and consistency properties of these estimators are obtained in Theorem 2 for the case where the criterion ξ_K is estimated by using empirical estimators as indicated in (8).

4.1 Estimation of σ and s

Let us consider a sequence $(f_n)_{n\in\mathbb{N}^*}$ of functions from I to $\mathbb{R}_+$ such that $f_n \sim n^{-\alpha} f$ where $\alpha \in]0, 1/2[$ and f is a strictly decreasing function from I to $\mathbb{R}_+$. Then, recalling that $K_i = I - \{i\}$, we put

$$\widehat{\phi}_i^{(n)} = \widehat{\xi}_{K_i}^{(n)} + f_n(i) \quad (i \in I) \tag{10}$$

and we take as estimator of σ the random permutation $\widehat{\sigma}^{(n)}$ of I such that

$$\widehat{\phi}_{\widehat{\sigma}^{(n)}(1)}^{(n)} \geq \widehat{\phi}_{\widehat{\sigma}^{(n)}(2)}^{(n)} \geq \cdots \geq \widehat{\phi}_{\widehat{\sigma}^{(n)}(p)}^{(n)}$$

and if $\widehat{\phi}^{(n)}_{\widehat{\sigma}^{(n)}(i)} = \widehat{\phi}^{(n)}_{\widehat{\sigma}^{(n)}(j)}$ with $i < j$, then the order is defined by $\widehat{\sigma}^{(n)}(i) < \widehat{\sigma}^{(n)}(j)$.

Remark 2 Just like σ, the permutation $\widehat{\sigma}^{(n)} = \left(\widehat{\sigma}^{(n)}(1), \ldots, \widehat{\sigma}^{(n)}(p)\right)$ is defined by ranking the $\widehat{\phi}^{(n)}_i$'s in nonincreasing order so that the ex aequo are ranked in increasing order of the corresponding indices. Then, $\widehat{\sigma}^{(n)}(i)$ is the index corresponding to the i-th largest element.

Furthermore, we consider the random set $\widehat{J}^{(n)}_i = \left\{\widehat{\sigma}^{(n)}(j)\,;\ 1 \le j \le i\right\}$ and the random variable

$$\widehat{\psi}^{(n)}_i = \widehat{\xi}^{(n)}_{\widehat{J}^{(n)}_i} + g_n\left(\widehat{\sigma}^{(n)}(i)\right) \qquad (i \in I) \tag{11}$$

where $(g_n)_{n\in\mathbb{N}^*}$ is a sequence of functions from I to $\mathbb{R}_+$ such that $g_n \sim n^{-\beta} g$ where $\beta \in]0, 1[$ and g is a strictly increasing function. Then, we take as estimator of s the random variable

$$\widehat{s}^{(n)} = \min\left\{i \in I \ / \ \widehat{\psi}^{(n)}_i = \min_{j\in I}\left(\widehat{\psi}^{(n)}_j\right)\right\}.$$

The variable selection is achieved by taking the random set

$$\widehat{I}^{(n)}_1 = \left\{\widehat{\sigma}^{(n)}(i)\ ;\ 1 \le i \le \widehat{s}^{(n)}\right\}$$

as estimator of I_1.

Remark 3 Since $\lim_{n\to+\infty} f_n(i) = 0$ and $\lim_{n\to+\infty} g_n(i) = 0$ for any $i \in I$, it is easily deduced from (10), (11) and Theorem 1 that $\widehat{\phi}^{(n)}_i$ and $\widehat{\psi}^{(n)}_i$ are consistent estimators of ξ_{K_i} and ξ_{J_i} (with $J_i := \{\sigma(j)\,;\ 1 \le j \le i\}$) respectively, just like $\widehat{\xi}^{(n)}_{K_i}$ and $\widehat{\xi}^{(n)}_{\widehat{J}^{(n)}_i}$. The penalty terms $f_n(i)$ and $g_n(\widehat{\sigma}^{(n)}(i))$ that are introduced in (10) and (11) permit to avoid ties in the values of $\widehat{\phi}^{(n)}_i$ and $\widehat{\psi}^{(n)}_i$. Indeed, even if one has $\widehat{\xi}^{(n)}_{K_i} = \widehat{\xi}^{(n)}_{K_j}$ (resp. $\widehat{\xi}^{(n)}_{J_i} = \widehat{\xi}^{(n)}_{J_j}$) for $i \ne j$, we will have $\widehat{\phi}^{(n)}_i \ne \widehat{\phi}^{(n)}_j$ (resp. $\widehat{\psi}^{(n)}_i \ne \widehat{\psi}^{(n)}_j$). This property is necessary, in the proof of Theorem 2 given below, for obtaining consistency of the proposed estimators $\widehat{\sigma}^{(n)}$ and $\widehat{s}^{(n)}$. If we directly use $\widehat{\xi}^{(n)}_{K_i}$ and $\widehat{\xi}^{(n)}_{\widehat{J}^{(n)}_i}$ consistency cannot be obtained because the aforementioned property may not be satisfied. This motivates the introduction of the penalty functions f_n and g_n in (10) and (11).

4.2 Consistency

When the empirical estimators defined in Sect. 3.1 are considerd, we establish consistency for the preceding estimators. We first give a proposition that is useful for proving the consistency theorem. There exist $t \in I$ and $(m_1, \ldots, m_t) \in I^t$ such that $m_1 + \cdots + m_t = p$, and $\xi_{K_{\sigma(1)}} = \cdots = \xi_{K_{\sigma(m_1)}} > \xi_{K_{\sigma(m_1+1)}} = \cdots = \xi_{K_{\sigma(m_1+m_2)}} > \cdots\cdots > \xi_{K_{\sigma(m_1+\cdots+m_{t-1}+1)}} = \cdots = \xi_{K_{\sigma(m_1+\cdots+m_t)}}$. We consider the set E of integers ℓ satisfying $1 \le \ell \le t$ and $m_\ell \ge 2$, and we put $m_0 := 0$ and

$F_\ell := \left\{\sum_{k=0}^{\ell-1} m_k + 1, \ldots, \sum_{k=0}^{\ell} m_k - 1\right\}$ $(\ell \in \{1, \ldots, t\})$. Then, introducing the assumption
$(\mathcal{A}_4)$: $\mathbb{E}\left(\|X\|^4\right) < +\infty$,
we have:

Proposition 2 *We assume that the asumptions* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_4)$ *hold, and that* $E \neq \varnothing$. *Then for any* $\ell \in E$ *and any* $i \in F_\ell$, $n^\alpha \left(\widehat{\xi}^{(n)}_{K_{\sigma(i)}} - \widehat{\xi}^{(n)}_{K_{\sigma(i+1)}}\right)$ *converges in probability to* 0, *as* $n \to +\infty$.

The following theorem gives consistency of the estimators $\widehat{\sigma}^{(n)}$ and $\widehat{s}^{(n)}$ defined in Sect. 4.1. The proof of this theorem is similar to that of Theorem 3.1 in Nkiet (2012).

Theorem 2 *We assume that the asumptions* $(\mathcal{A}_1)$, $(\mathcal{A}_2)$ *and* $(\mathcal{A}_4)$ *hold. Then, we have:*

(i) $\lim_{n\to+\infty} P\left(\widehat{\sigma}^{(n)} = \sigma\right) = 1$;
(ii) $\widehat{s}^{(n)}$ *converges in probability to* s, *as* $n \to +\infty$.

As a consequence of this theorem, we easily obtain:

$$\lim_{n\to+\infty} P\left(\widehat{I}_1^{(n)} = I_1\right) = 1.$$

This shows the consistency of our method for selecting variables in discriminant analysis with mixed variables.

Remark 4 Technical arguments in the proofs of Proposition 2 and Theorem 2 motivate the introduction of f_n and g_n in (10) and (11) (see Remark 3). They also explain the choice of f, g, α and β with the related properties. Indeed:

(i) in the proof of Proposition 2 we have, for instance, the inequality

$$|\widehat{A}_i^{(n)}| \leq n^{\alpha-1/2}\left(\|\widehat{\Lambda}^{(n)}_{K_{\sigma(i)}|m}\|_\infty + \|\widehat{\Lambda}^{(n)}_{K_{\sigma(i+1)}|m}\|_\infty\right)\|\widehat{W}^{(n)}\|_{\mathcal{E}}$$

from which we want to prove that $\widehat{A}_i^{(n)}$ converges in probability to 0 as $n \to +\infty$. Since, as $n \to +\infty$, $\|\widehat{\Lambda}^{(n)}_{K_{\sigma(i)}|m}\|_\infty + \|\widehat{\Lambda}^{(n)}_{K_{\sigma(i+1)}|m}\|_\infty$ converges almost surely to $\|\Lambda_{K_{\sigma(i)}|m}\|_\infty + \|\Lambda_{K_{\sigma(i+1)}|m}\|_\infty$ and $\|\widehat{W}^{(n)}\|_{\mathcal{E}}$ converges in distribution to $\|W\|_{\mathcal{E}}$, where W has a normal distribution, we have to take $\alpha < 1/2$ in order to obtain the required convergence property. In addition, for having $\lim_{n\to+\infty} f_n(i) = 0$ we must take $\alpha > 0$.

(ii) In the proof of Theorem 2, a similar argument leads to $0 < \beta < 1$. Further, we want to obtain $f(\sigma(i)) - f(\sigma(i+1)) > 0$ and $g(\sigma(i)) - g(\sigma(s)) > 0$ for $(i, s) \in I^2$ satisfying $\sigma(i) < \sigma(i+1)$ and $\sigma(i) > \sigma(s)$. That is why f (resp. g) is taken as a strictly decreasing (resp. increasing) function.

Remark 5 The variable selection method that is described above can be performed by using smoothed non-parametric estimators defined in Sect. 3.2. It suffices to replace $\widehat{\xi}^{(n)}_{K_i}$ and $\widehat{\xi}^{(n)}_{\widehat{J}_i}$ by $\widetilde{\xi}^{(n)}_{K_i}$ and $\widetilde{\xi}^{(n)}_{\widehat{J}_i}$ in (10) and (11). But, since no consistency results are known for these estimators, we could not garantee consistency of the resulting method.

5 Numerical experiments

In this section, we report results of simulations made for studying properties of the proposed method. Several issues are adressed: the influence of the penalty functions f_n and g_n introduced in (10) and (11), the type of estimator and the parameters α and β on the performance of the procedure, optimal choice of these parameters and comparison with the method proposed in Mahat et al. (2007).

5.1 The simulated data sets

Each data set was generated as follows: for a given value of p, X_i is generated from a multivariate normal distribution in $\mathbb{R}^p$ with mean μ and covariance matrix given by $\Gamma = \frac{1}{2}(\mathbb{I}_p + J_p)$, where $\mathbb{I}_p$ is the $p \times p$ identity matrix and J_p is the $p \times p$ matrix whose elements are all equal to 1. For a given d and $M = 2^d$, U_i is generated from a discrete distribution on $\{1, \ldots, M\}$ with probabilities $q_1, \ldots, q_M$ such that $\sum_{k=1}^{M} q_k = 1$, that is equivalent to generate Y_i as random vector with d coordinates being binary random variables. For this, two models are used:

(i) **Model 1:** $q_1 = q_2 = \cdots = q_M = 1/M$ (uniform distribution);
(ii) **Model 2:** $M = 8$ and $q_1 = q_3 = q_5 = 0.001$, $q_2 = q_4 = q_6 = q_7 = q_8 = 0.1994$.

Model 2 corresponds to the case where many cell incidences are very small and will be relevant in order to compare empirical and non-parametric estimators. Two groups of data was generated as indicated above with $\mu = \mu_1 = (0, \ldots, 0)^T$ for the first group and $\mu = \mu_2 = (\mu_1^{(2)}, \ldots, \mu_p^{(2)})^T$ for the second group, where

$$\mu_{2k}^{(2)} = 0 \quad \text{and} \quad \mu_{2k-1}^{(2)} = k/p$$

for $k = 1, \ldots, [(p+1)/2]$, the notation $[x]$ denoting the integer part of x. Our simulated data is based on two independent data sets: training data and test data, each with sample size $n = 100, 300, 500, 1000$ and with size $n_1 = n_2 = n/2$ for the two groups. The training data is used for selecting variables and the test data is used for computing the correct classification rate (CCR), that is the proportion of correct classification. For the two groups case (i.e. when $q = 2$) classification after variable selection is achieved by using the rule (1) assuming equal costs and equal prior probabilities in both two groups (hence $\log(\gamma) = 0$), and for the multiple classes case (i.e. when $q > 2$) the rule (2) is used assuming that $\tau_1 = \cdots = \tau_q$. The average of CCR over 1000 independent replications is used for measuring the performance of the methods.

5.2 Influence of penalty functions and type of estimator

In order to evaluate the impact of penalty functions on the performance of our method we took $f_n(i) = n^{-\alpha}/h_k(i)$ and $g_n(i) = n^{-\beta}h_k(i), k = 1, \ldots, 13$, with $\alpha = \beta = 1/4$, $h_1(x) = x$, $h_2(x) = x^{0.1}$, $h_3(x) = x^{0.5}$, $h_4(x) = x^{0.9}$, $h_5(x) = x^{10}$, $h_6(x) =$

$\ln(x)$, $h_7(x) = \ln(x)^{0.1}$, $h_8(x) = \ln(x)^{0.5}$, $h_9(x) = \ln(x)^{0.9}$, $h_{10}(x) = x\ln(x)$, $h_{11}(x) = (x\ln(x))^{0.1}$, $h_{12}(x) = (x\ln(x))^{0.5}$, $h_{13}(x) = (x\ln(x))^{0.9}$. For each of these functions, we computed CCR by using both empirical estimators from Sect. 3.1 and non-parametric smoothing procedure introduced in Sect. 3.2. For this latter type of estimator, the smoothing parameter λ was computed from a cross validation method on the training sample in order to maximize correct classification rate. The results for Model 1 are given in Table 1. It is observed that there is no significant difference between the results obtained for the different functions. So, it seems that choosing penalty functions has no influence on the performance of our method. Also, the results in Table 1 don't really show any significant difference between the empirical and smoothed estimators when cell entries of the d-dimensional variate contingency table corresponding to the d categorical variables are not small. However, when Model 2 is used, i.e in the case where the aforementioned cell entries are very small, we see in Table 2 that estimates based on empirical estimators can often not be computed (NA=not available), certainly because of the absence of observations in some cells, while the results are suitable with smoothed estimators. This suggests that one should prefer to use smoothed estimator for performing the proposed method.

5.3 Influence of parameters α and β

Since tuning parameters may have impact on the performance of a statistical procedure, it is important to study their influence. That is why numerical experiments have been carried out in order to investigate the influence of α and β on the performance of our method. For doing that, we made simulations as indicated above by taking

$$f_n(i) = n^{-\alpha}\ln(i)^{-0.1} \quad \text{and} \quad g_n(i) = n^{-\beta}\ln(i)^{0.1} \tag{12}$$

with $\alpha = 0.1, 0.2, 0.3, 0.4, 0.45$, and β varying in $[0, 1[$. The results are reported in Fig. 1a–c. Although the obtained curves vary as α and β vary, we cannot say that these parameters have a marked impact on the performance of our method. Indeed, all differences in CCR values are negligible since they do not exceed 0.003. Thus, one can think that, in practice, our method can be performed with arbitrary values for α and β without significantly affecting its performance. However, an interesting alternative would be to determine optimal values of these parameters. An approach for doing that via cross-model validation is described in the following section.

5.4 Choosing optimal (α, β)

We propose a method for making an optimal choice of (α, β) based on leave-one-out cross validation used in order to maximize corect classification rate. For each $k \in \{1, \ldots, n\}$, after removing the k-th observation for X and Y in the training sample our method for selecting variable is applied on this remaining sample with a given value for (α, β) and penalty functions taken as in (12). Then, the observation that have been removed is allocated to a group $\widetilde{g}_{\alpha,\beta}(k)$ in $\{1, \ldots, q\}$ by using the rule given in

Table 1 Average of CCR values over 1000 replications for Model 1 with $q = 2$ classes, $d = 3$ binary variables, $M = 2^3 = 8$, and $p = 5$ continuous variables

Function	CCR	
	Empirical estimators	Non-parametric estimators
$n = 100\ (n_1 = n_2 = 50)$		
h_1	0.60800	0.60800
h_2	0.60900	0.60900
h_3	0.61000	0.61000
h_4	0.60900	0.60900
h_5	0.60800	0.60800
h_6	0.61028	0.61028
h_7	0.60903	0.60903
h_8	0.61066	0.61066
h_9	0.60898	0.60898
h_{10}	0.60842	0.60842
h_{11}	0.60948	0.60948
h_{12}	0.60915	0.60915
h_{13}	0.60908	0.60908
$n = 300\ (n_1 = n_2 = 150)$		
h_1	0.56421	0.56424
h_2	0.56328	0.56332
h_3	0.56417	0.56417
h_4	0.56346	0.56346
h_5	0.56382	0.56382
h_6	0.56343	0.56343
h_7	0.56374	0.56374
h_8	0.56354	0.56354
h_9	0.56312	0.56312
h_{10}	0.56379	0.56379
h_{11}	0.56404	0.56404
h_{12}	0.56345	0.56345
h_{13}	0.56375	0.56375
$n = 500 (n_1 = n_2 = 250)$		
h_1	0.54998	0.54998
h_2	0.55047	0.55067
h_3	0.55032	0.55039
h_4	0.55049	0.55058
h_5	0.54982	0.55982
h_6	0.55058	0.55062
h_7	0.55050	0.55054
h_8	0.55008	0.55013
h_9	0.54972	0.54934

Table 1 continued

Function	CCR	
	Empirical estimators	Non-parametric estimators
h_{10}	0.55013	0.55018
h_{11}	0.55036	0.55037
h_{12}	0.55046	0.55046
h_{13}	0.55028	0.55028

Weightings obtained from 13 penalty functions $f_n = n^{-1/4}/h_k$ and $g_n = n^{-1/4}h_k$, $k = 1, \ldots, 13$

Table 2 Average of CCR values over 1000 replications for Model 2 with $q = 2$ classes, $d = 3$ binary variables, $M = 2^3 = 8$, and $p = 5$ continuous variables

n	$n_1 = n_2$	CCR	
		Empirical estimators	Non-parametric estimators
100	50	NA	0.54000
300	150	NA	0.56670
500	250	NA	0.52400

Weightings obtained from penalty functions $f_7 = n^{-1/4}/h_7$ and $g_7 = n^{-1/4}h_7$

(1) (for the two groups case) or in (2) (for the case of more than two groups) based on the variables that have been selected in the previous step. Then, we consider

$$CV(\alpha, \beta) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{Z_k = \widetilde{g}_{\alpha,\beta}(k)\}}$$

and we take as optimal value for (α, β) the pair $(\alpha_{opt}, \beta_{opt})$ defined by:

$$(\alpha_{opt}, \beta_{opt}) = \underset{(\alpha,\beta)\in]0,1/2[\times]0,1[}{\operatorname{argmax}} CV(\alpha, \beta). \tag{13}$$

5.5 Algorithm

Algorithm1 describes the proposed method for practically choose optimal (α, β) from simulated data. It can obviously be adapted for permorming our method on real data sets.

5.6 Comparison with the method of Mahat et al. (2007)

In order to compare our method to that of Mahat et al. (2007), 1000 independent replications were made. For each of these replications the method described in Algorithm 1 was performed:

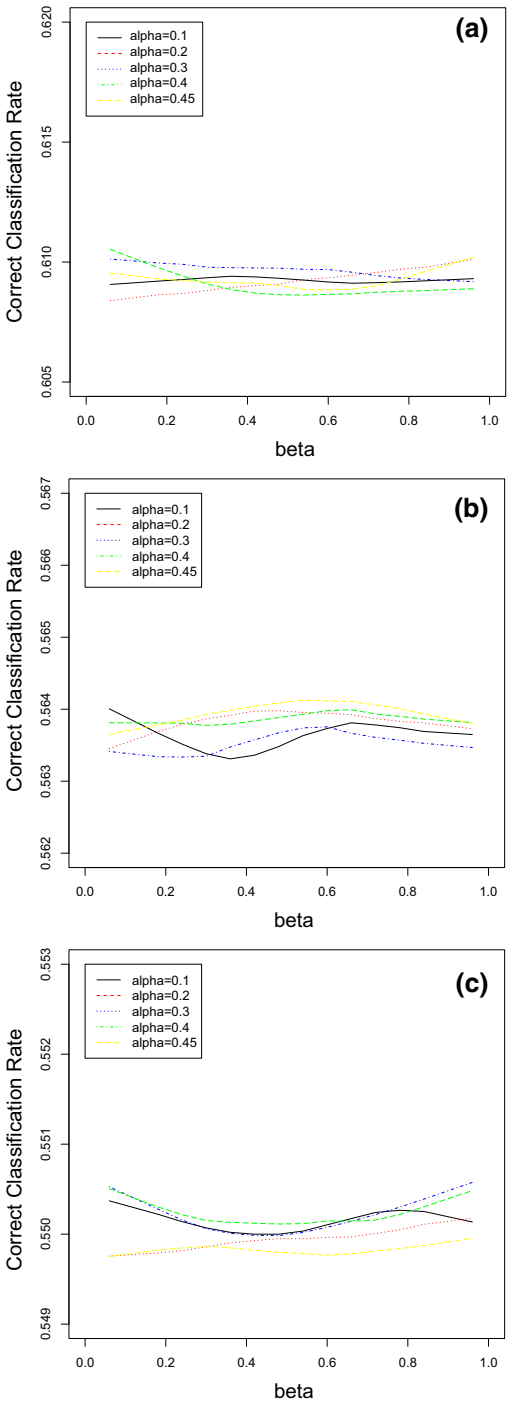

Fig. 1 Average of CCR values over 1000 replications for Model 1 with $q = 2$ classes, $d = 3$ binary variables, $M = 2^3 = 8$, $p = 5$ continuous variables, $\alpha = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.45$ and β varying in $]0, 1[$. Weightings obtained from penalty functions $f_7 = n^{-\alpha}/h_7$ and $g_7 = n^{-\beta}h_7$. **a** $n = 100$. **b** $n = 300$. **c** $n = 500$

Algorithm 1 Computation of Correct Classification Rate with simulated data

Simulate a training data set $\mathcal{S}_1$ as indicated in 5.1

Finely discretize the rectangle $R =]0, 1/2[\times]0, 1[$ so as to obtain a grid of points G

for all $(\alpha, \beta) \in G$ **do**

for $k = 1, \cdots, n$ **do**

Remove the k-th observation from this training data set, the remaining data set is denote by $\mathcal{S}_1^{(-k)}$

Apply variable selection method on $\mathcal{S}_1^{(-k)}$ so as to obtain a set $\widehat{I}_1^{(-k)}$ of relevant variables

Allocate the removed observation to one of the groups by using the rules (1) or (2), the variables selected in $\widehat{I}_1^{(-k)}$, and estimates of parameters involved in the rules from the data in $\mathcal{S}_1^{(-k)}$

Letting $\widetilde{g}_{\alpha,\beta}(k)$ be the group to which the removed observation is allocated in the previous step, set $\epsilon_k = 1$ if $\widetilde{g}_{\alpha,\beta}(k)$ is the real group of this observation, and $\epsilon_k = 0$ otherwise

end for

Set $CV(\alpha, \beta) = n^{-1} \sum_{k=1}^{n} \epsilon_k$

end for

Take $(\widehat{\alpha}, \widehat{\beta})$ that maximizes $CV(\alpha, \beta)$ over G

Apply variable selection method on the whole training set $\mathcal{S}_1$ with $(\widehat{\alpha}, \widehat{\beta})$ so as to obtain a set $\widehat{I}_1$ of relevant variables

Simulate an independent test data set $\mathcal{S}_2$ as indicated in 5.1

Compute the correct classification rate (CCR) on $\mathcal{S}_2$ by using the variables selected in $\widehat{I}_1$

(i) the training sample is used for selecting variables from our method and that of Mahat et al. (2007); our method is used with penalty functions given in (12) and optimal (α, β) obtained by using leave-one-out cross validation as indicated in (13);
(ii) the test sample is then used for computing CCR for the two methods.

Model 1 was used with large number of variables: the number of continuous variables is $p = 100$, and the number of binary variables is $d = 6, 8$; then, the number of cells of the resulting multinomial variable is $M = 64{,}256$. The average of CCR over the 1000 replications is then computed. Table 3 gives the results for three methods:

- our method with empirical estimators (Meth1);
- our method with smoothed non-parametric estimators (Meth2);
- the method of Mahat et al. (2007) (Meth3).

The average of CCR before variable selection (denoted by CCRb in Table 3 and Table 4) is also computed in order to compare the results with that corresponding to the 'true' model involving all p continuous variables. The results in Table 3 do not show a superiority of one of the methods compared to the other. However, they give

Table 3 Average of CCR values over 1000 replications for Model 1 with $q = 2$ classes, $d = 6, 8$ binary variables, $M = 64, 256$, and $p = 100$ continuous variables

n	$n_1 = n_2$	$d = 6$				$d = 8$			
		Meth1	Meth2	Meth3	CCRb	Meth1	Meth2	Meth3	CCRb
300	150	NA	0.6300	0.6300	0.7234	NA	0.7200	0.7200	0.8754
500	250	NA	0.6375	0.6370	0.6697	NA	0.7000	0.7000	0.8208
1000	500	NA	0.5800	0.5800	0.6118	NA	0.6700	0.6700	0.7444

Weightings obtained from penalty functions $f_n = n^{-\alpha_{opt}}/h_k$ and $g_n = n^{-\beta_{opt}} h_k$, where $(\alpha_{opt}, \beta_{opt})$ is obtained from cross-validation

Table 4 Average of CCR values over 1000 replications for Model 1 with $q = 3$ classes, $d = 3, 8$ binary variables, $M = 8, 256$, and $p = 25$ continuous variables

n	$n_1 = n_2 = n_3$	$d = 3$			$d = 8$		
		Meth1	Meth2	CCRb	Meth1	Meth2	CCRb
600	200	0.3840	0.4900	0.4495	NA	0.3830	0.6653
900	300	0.4530	0.5730	0.4349	NA	0.4820	0.6649
1200	400	0.4720	0.5470	0.4252	NA	0.5150	0.6489

Weightings obtained from penalty functions $f_n = n^{-\alpha_{opt}}/h_k$ and $g_n = n^{-\beta_{opt}} h_k$, where $(\alpha_{opt}, \beta_{opt})$ is obtained from cross-validation

another illustration of the interest to use our method with smoothed non-parametric estimators rather than empirical estimators, especially in high-dimensional context. Indeed, when $d = 8$ and $n = 300, 500$, there are more cells than observations in each dataset. Therefore, some cells do not have any incidence. When $d = 6$ or when $d = 8$ and $n = 1000$, some cells may have no incidence. Then estimates based on empirical estimators cannot be computed while this problem does not arise for the smoothed non-parametric estimators. Comparing the results with that obtained from the 'true' model, it appears that our method has good behavior since the difference between CCRb and CCR is not large, especially for large values of n.

One of the advantages of the method we propose is that it can be used when there are more than two groups, which is not the case for classical methods, especially that of Mahat et al. (2007). For illustrating this fact, we made simulations for the case of three groups. Model 1 was used again with the first two groups taken as above, and a third group corresponding to the mean $\mu_3 = (\mu_1^{(3)}, \ldots, \mu_p^{(3)})^T$ with:

$$\mu_{2k}^{(3)} = k/p \quad \text{and} \quad \mu_{2k-1}^{(3)} = 0,$$

for $k = 1, \ldots, [(p+1)/2]$. Taking $p = 25$ and $d = 3, 8$ we computed the average of CCR over 1000 replications. The results are reported in Table 4. We obtained better results with smoothed non-parametric estimators. When $d = 3$ the obtained results are better than these from the 'true' model, and when $d = 8$ the difference between CCRb and CCR is larger.

6 A data example

To further demonstrate its practical usefulness, we apply our method to the cover type dataset from the Repository of Machine Learning Databases maintained by the University of California at Irvine. This dataset consists of 1818 tree observations from areas of the Roosevelt National Forest in Colorado. The units are the 1818 trees for which several cartographic variables are observed. There are seven tree types, each represented by an integer variable: (i) Spruce/Fir; (ii) Lodgepole Pine; (iii) Ponderosa Pine; (iv) Cottonwood/Willow; (v) Aspen; (vi) Douglas-fir; (vii) Krummholz. The tree type is an observed variable that induces 7 classes of units. For each unit 10 numerical variables and 44 binary variables are observed. The numerical variables are: (V1) elevation (in meters); (V2) aspect (in degrees azimuth); (V3) slope (in degrees); (V4) horizontal distance to nearest surface water features; (V5) vertical distance to nearest surface water features; (V6) horizontal distance to nearest roadway; (V7) hillshade index at 9am; (V8) hillshade index at noon; (V9) hillshade index at 3pm; (V10) horizontal distance to nearest wildfire ignition points. We considered 10 of the 44 aforementioned binary variables: 4 binary variables giving wilderness area designation, and 6 binary variables defining soil type.

This data set is suitable for classification as it permits to predict the type of a given tree for which the above binary and numerical variables are observed, by using an allocation rule like the one defined in (2). We are interested in determining, among the 10 numerical variables given above, those that are relevant for this classification approach. For doing that, we applied our method on this data set, with $n = 1818$, $p = 10$, $q = 7$, $d = 10$ and $M = 2^{10} = 1024$, and by using the smoothed non-parametric estimators defined in 3.2. The penalty functions used were the function given in (12). The data set was divided in two parts of equal sizes $n_1 = n_2 = n/2$. The first part was used as a training set for estimating optimal α and β via leave-one-out cross validation as indicated in Sect. 5.4, and for selecting variables i.e. estimating the set of relevant numerical variables for discriminating between the seven groups. The second part is then used as a test sample for computing CCR afer variable selection by using the rule (2) assuming that $\tau_1 = \cdots = \tau_7$. We found that the relevant numerical predictors are $V1$, $V6$ and $V10$, and that $CCR = 0.5830$. We also computed CCR with all 10 continuous variables (denoted by CCRb) in order to compare it with the preceding value, and we obtained $CCRb = 0.5940$. These results suggests that:

- elevation (V1), horizontal distance to nearest roadway (V6) and horizontal distance to nearest wildfire ignition points (V10) are the most relevant variables for predicting cover type in the presence of the binary variables giving wilderness area designation and defining soil type;
- our method behaves well since the difference between CCRb and CCR is quite small.

This example also shows the ability of our method to perform in case there are more than two groups (here, there are $q = 7$ groups) and when both continuous and binary variable have high dimensions.

7 Conclusion

In this paper, we introduce a variable selection method for discrimination among several groups with both continuous and categorical predictors. The main advantages of the proposal are: (i) no assumption on the distribution of the involved predictors is needed, in particular the location model, which is classically assumed in this context, is not supposed to hold; (ii) it can be used for more than two groups, which is not the case for conventional methods which are limited to the case of two groups. Both theoretical results and numerical studies show that our method behaves well, especially when smoothed non-parametric estimators are used for estimating the proposed criterion.

Acknowledgements We are very grateful to two anonymous referees for their helpful and constructive comments, which led to a much improved manuscript. Research by Alban Mbina Mbina was supported in part by the Agence Universitaire de la Francophonie (AUF).

References

Aspakourov O, Krzanowski WJ (2000) Non-parametric smoothing of the location model in mixed variables discrimination. Stat Comput 10:289–297

Bar-Hen A, Daudin JJ (1995) Generalization of the Mahalanobis distance in the mixed case. J Multivar Anal 53:332–342

Bedrick EJ, Lapidus J, Powell JF (2000) Estimating the Mahalanobis distance from mixed continuous and discrete data. Biometrics 56:394–401

Chang PC, Afifi AA (1979) Classification based on dichotomous and continue variables. J Am Stat Assoc 69:336–339

Daudin JJ (1986) Selection of variables in mixed-variable discriminant analysis. Biometrics 42:473–481

Daudin JJ, Bar-Hen A (1999) Selection in discriminant analysis with continuous and discrete variables. Comput Stat Data Anal 32:161–175

De Leon AR, Carriere KC (2005) A generalized Mahalanobis distance for mixed data. J Multivar Anal 92:174–185

De Leon AR, Soo A, Williamson T (2011) Classification with discrete and continuous variables via general mixed-data models. J Appl Stat 38:1021–1032

Fujikoshi Y (1982) A test for additional information in canonical correlation analysis. Ann Inst Stat Math 34:523–530

Fujikoshi Y (1985) Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. J Multivar Anal 17:27–37

Hand DJ (1997) Construction and assessment of classification rules. Wiley, Chichester

Krusinska E (1989a) New procedure for selection of variables in location model for mixed variable discrimination. Biom J 31:511–523

Krusinska E (1989b) Two step semi-optimal branch and bound algorithm for feature selection in mixed variable discrimination. Pattern Recognit. 22:455–459

Krusinska E (1990) Suitable location model selection in the terminology of graphical models. Biom J 32:817–826

Krzanowski WJ (1975) Discrimination and classification using both binary and continuous variables. J Am Stat Assoc 70:782–790

Krzanowski WJ (1983) Stepwise location model choice in mixed variable discrimination. J R Stat Soc C 32:260–266

Krzanowski WJ (1984) On the null distribution of distance between two groups, using mixed continuous and categorical variables. J Classif 1:243–253

Mahat NI, Krzanowski WJ, Hernandez A (2007) Variable selection in discriminant analysis based on the location model for mixed variables. Adv Data Anal Classif 1:105–122

McKay RJ (1977) Simultaneous procedures for variable selection in multiple discriminant analysis. Biometrika 64:283–290

McLachlan GJ (1992) Discriminant analysis and statistical pattern recognition. Wiley, New York
Nkiet GM (2012) Direct variable selection for discrimination among several groups. J Multivar Anal 105:151–163
Olkin I, Tate RF (1961) Multivariate correlation models with mixed discrete and continuous variables. Ann Math Stat 32:448–465. J Multivar Anal 105:151–163